Data Mining: Classification

Jakramate Bootkrajang

School of Computer Science
Chiang Mai University

Adapted from materials by Ata Kabán

Outline

- Introduction
- Bayesian Learning
- Generative Classifier
- Discriminative Classifier
- Nearest Neighbour Classifier
- Classifier Evaluation [Maybe]

Intro: The three learning paradigms

- Supervised learning (Classification)
  - The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
  - New data is classified based on the training set
- Unsupervised learning (Clustering)
  - No label as the class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of finding the existence of classes or clusters in the data
- Reinforcement learning (Robot training)

Typical applications of classification

- Credit/loan approval
- Medical diagnosis: if a tumour is cancerous or benign
- Fraud detection: if a transaction is fraudulent
- Web page categorisation: which category it is ?

What will we learn in this lecture?

- We will learn about
  - Bayes rule
  - Turn this into a classifier
- Scenario: Build a classifier for predicting if a patient is ill or healthy based on
  - A probabilistic model of the observed data
  - Prior knowledge (important in Bayesian framework)

Classification problem

- Training data: Example of the form $(x, h(x))$.
  - x are the data objects to classify (inputs)
  - $h(x)$ are the correct class info for x, $h(x) \in \{1, \ldots, K\}$.

- Goal: Given x_{new}, we wants to know $h(x_{new})$.
- Error: $|\text{target output} - \text{predicted output}|$

A word about the Bayesian Framework

- Allows us to combine observed data and prior knowledge
- Provides practical learning algorithms
- It is a generative (model-based) approach, which offers a useful conceptual framework
  - This means that any kind of objects (e.g. time-series, trees, etc.) can be classified, based on a probabilistic model specification

Bayes' Rule

- $P(h|D) = \frac{P(D|h)P(h)}{P(D)}$
- Who is who in Bayes' rule
  - h = hypothesis, D = dataset, x = data point
  - $P(h)$: prior belief (probability of hypothesis)
  - $P(D|h)$: likelihood (probability of the data if the hypothesis h is *true*)
  - $P(D) = \sum_h P(D|h)P(h)$: Data evidence (marginal probability of the data)
  - $P(h|D)$ = posterior (probability of hypothesis h after having seen the data D)
  - The numerator is essentially the joint prob. $P(D|h)P(h) = P(D, h)$

A Side Note on Probability

- Suppose we have two dices h_1 and h_2
  - Say, h_1 is fair but h_2 is biased
- The probability of getting i given the h_1 dice is denoted $P(i|h_1) \rightarrow$ conditional probability
- Pick a dice at random with $P(h_j) : j = 1, 2$. The probability for picking the h_j dice *and* getting an i with the dice, is called joint probability and is $P(i, h_j) = P(h_j)P(i|h_j)$

A Side Note on Probability

- For events X, Y: $P(X, Y) = P(X|Y)P(Y)$.
- If we know $P(X, Y)$, then the so-called marginal probability is $P(X) = \sum_Y P(X, Y)$.

Does a patient have cancer or not?

- A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 98% of the cases and a correct negative result in only 97% of the cases.
  1. What's the probability that this patient has cancer?
  2. What's the probability that he does not have cancer?
  3. What is the diagnosis?

Choosing hypotheses

- Want: $h = \arg\max_{h \in H} P(h|D)$
- Two general ways to do this
  1. Maximum Likelihood (does not consider priors, assume equal priors):
     $h_{ML} = \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)$
  2. Maximum a Posteriori (consider prior):
     $h_{MAP} = \arg\max_{h \in H} P(h|D)$

ML Solution

- h_1 = 'cancer', h_2 = '$\neg$ cancer', $data$ = '+'

1. $P(cancer|+) = P(+|cancer) = 0.98$
2. $P(\neg cancer|+) = 1 - 0.98 = 0.02$
3. Diagnosis ? He has cancer.

Does a patient have cancer or not? (new info)

- A patient takes a lab test and the result comes back positive. It is known that the test returns a correct positive result in only 98% of the cases and a correct negative result in only 97% of the cases. *Furthermore, only 0.008 of the entire population has this disease.*
  1. What's the probability that this patient has cancer?
  2. What's the probability that he does not have cancer?
  3. What is the diagnosis?

MAP Solution

- $h_1 =$ 'cancer', $h_2 =$ '$\neg$ cancer', $data =$ '+'

1. $P(cancer|+) = \frac{P(+|cancer)P(cancer)}{P(+)} = \frac{0.98 \times 0.008}{xx} = 0.21$

   $P(+|cancer) = 0.98,\ P(cancer) = 0.008$

   $P(+) = \sum_j P(+|h_j)P(h_j)$
   $P(+|cancer)P(cancer) + P(+|\neg cancer)P(\neg cancer)$

   $P(+|\neg cancer) = 0.03$

   $P(cancer)$ = 1-0.008

2. $P(\neg cancer|+) = 1 - 0.21$

3. Diagnosis ? Negative.

The Naive Bayes Classifier (1/2)

- What can we do if our data d has several attributes? $\mathbf{x} = \{a_1, a_2, \ldots, a_m\}$
- The problem is $P(h, \mathbf{x}) = P(h)P(\mathbf{x}|h)$ factorised into a long sequence.
- By chain rule $P(h, \mathbf{x}) = P(h)P(a_1, \ldots, a_m|h)$

  $= P(h)P(a_1|h)P(a_2, \ldots, a_m|h, a_1)$

  $= P(h)P(a_1|h)P(a_2|h, a_1)P(a_3, \ldots, a_m|h, a_1, a_2)$
- The naive assumption assumes that each feature a_i is conditionally independent of every other feature a_j for $j \neq i$

The Naive Bayes Classifier (2/2)

- So we have $P(a_i|h, a_j) = P(a_i|h)$, $P(a_i|h, a_j, a_k) = P(a_i|h)$ and so on.
- Which gives: $P(\mathbf{x}|h) = P(a_1, \ldots, a_m) = \prod_i P(a_i|h)$
- A Bayesian classifier that uses the Naive assumption is called The Naive Bayes classifier.
- One of the most practical methods
- Widely used in
  - medical applications
  - text classification.

Example of Naive classifier: Playing Tennis

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
Day1	Sunny	Hot	High	Weak	*No*
Day2	Sunny	Hot	High	Strong	*No*
Day3	Overcast	Hot	High	Weak	Yes
Day4	Rain	Mild	High	Weak	Yes
Day5	Rain	Cool	Normal	Weak	Yes
Day6	Rain	Cool	Normal	Strong	*No*
Day7	Overcast	Cool	Normal	Strong	Yes
Day8	Sunny	Mild	High	Weak	*No*
Day9	Sunny	Cool	Normal	Weak	Yes
Day10	Rain	Mild	Normal	Weak	Yes
Day11	Sunny	Mild	Normal	Strong	Yes
Day12	Overcast	Mild	High	Strong	Yes
Day13	Overcast	Hot	Normal	Weak	Yes
Day14	Rain	Mild	High	Strong	*No*

Naive Bayes Solution

- Classify any new data point $\mathbf{x} = (a_1, \ldots, a_m)$ as
- $h_{naive} = \arg\max_h P(h)P(\mathbf{x}|h) = \arg\max_h P(h)\prod_i P(a_i|h)$
- We need to estimate the parameters from the training examples
  - For each hypothesis h: $\hat{P}(h) :=$ estimated$P(h)$
  - For each feature a_i: $\hat{P}(a_i|h) :=$ estimated$P(a_i|h)$
- Based on the examples in the table, classify the following $\mathbf{x}$.
  $\mathbf{x} = \{Outlook = Sunny, Temp = Cool, Hum = High, Wind = Strong\}$ Play tennis or not ?

The working

- $h_{naive} = \arg\max_{h \in [yes,no]} P(h)P(\mathbf{x}|h)$

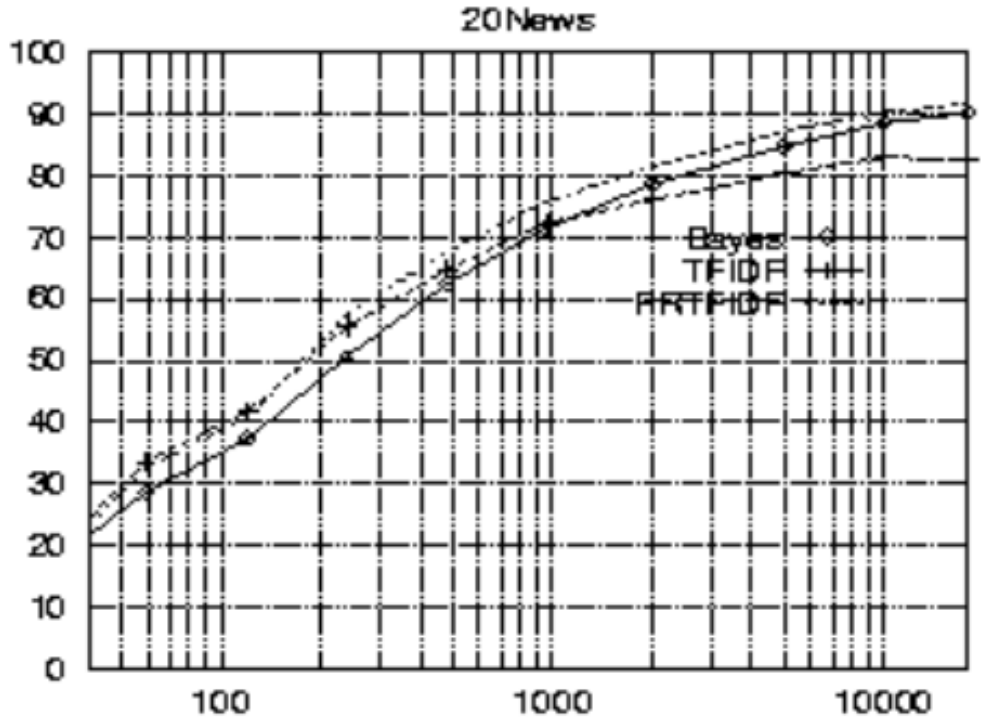
  $= \arg\max_{h \in [yes,no]} P(h) \prod_i P(a_i|h)$

  $= \arg\max P(h)P(Outlook = sunny|h)P(Temp = cool|h)P(Humidity = high|h)P(Wind = strong|h)$
- Now find $P(Playtennis = Yes) = 9/14 = 0.64$
- find $P(Playtennis = No) = 5/14 = 0.36$
- find $P(Wind = strong|Playtennis = Yes) = 3/9 = 0.33$
- find $P(Wind = strong|Playtennis = No) = 3/5 = 0.60$
- And so on. Finally we'll see that we don't play tennis today

Learning to classify text

- Learn from examples which articles are of interest
- The attributes (features) are the words
- NB classifiers are one of the most effective for this task

Example of 20-Newsgroups text classification using NB



Accuracy vs. Training set size (1/3 withheld for test)

Summary

- Bayes' rule can be turned into a classifier
- Maximum A Posteriori (MAP) hypothesis estimation incorporates prior knowledge; Maximum Likelihood doesn't
- Naive Bayes classifier is a simple but effective Bayesian classifier for vector data (i.e. data with several attributes) that assumes that attr. are independent given the class.
- Bayesian classification is a generative approach to classification.